PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification [6] :
G10L

A2

(11) International Publication Number: WO 98/35340

(43) International Publication Date: 13 August 1998 (13.08.98)

(21) International Application Number: PCT/US98/01538

(22) International Filing Date: 27 January 1998 (27.01.98)

(30) Priority Data:
60/036,227 27 January 1997 (27.01.97) US

(71) Applicant *(for all designated States except US)*: ENTROPIC RESEARCH LABORATORY, INC. [US/US]; Suite G100, 400 North Capitol Street, N.W., Washington, DC 20001 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: ARSLAN, Levent, M. [TR/US]; 850 Randolph Street #811, Arlington, VA 22203 (US). TALKIN, David [US/US]; 1727 Lansing Court, McLean, VI 22101 (US).

(74) Agents: CARLSON, Stephen, C. et al.; Lowe Price Leblanc & Becker, Suite 300, 99 Canal Center Plaza, Alexandria, VA 22314 (US).

(81) Designated States: AU, CA, IL, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: VOICE CONVERSION SYSTEM AND METHODOLOGY

(57) Abstract

A voice conversion system and methodology employ a codebook mapping approach to transforming a source voice to sound like a target voice. Each speech frame is represented by a weighted average of codebook entries. The weights represent a perceptual distance of the speech frame and may be refined by a gradient descent analysis. The vocal tract characteristics, represented by a line spectral frequency vector, the excitation characteristics, represented by a linear predictive coding residual, the duration, and the amplitude of the speech frame are transformed in the same weighted-average framework.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

VOICE CONVERSION SYSTEM AND METHODOLOGY

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application No. 60/036,227, entitled "Voice Conversion by Segmental Codebook Mapping of Line Spectral Frequencies and Excitation System," filed on January 27, 1997 by Levent M. Arslan and David Talkin, incorporated herein by reference.

FIELD OF THE INVENTION

The present invention relates to voice conversion and, more particularly, to codebook-based voice conversion systems and methodologies.

BACKGROUND OF THE INVENTION

A voice conversion system receives speech from one speaker and transforms the speech to sound like the speech of another speaker. Voice conversion is useful in a variety of applications. For example, a voice recognition system may be trained to recognize a specific person's voice or a normalized composite of voices. Voice conversion as a front-end to the voice recognition system allows a new person to effectively utilize the system by converting the new person's voice into the voice that the voice recognition system is adapted to recognize. As a post processing step, voice conversion changes the voice of a text-to-speech synthesizer. Voice conversion also has applications in voice disguising, dialect modification, foreign-language dubbing to retain the voice of an original actor, and novelty systems such as celebrity voice impersonation, for example, in Karaoke machines.

In order to convert speech from a "source" voice to a "target" voice, codebooks of the source voice and target voice are typically prepared in a training phase. A codebook is a collection of "phones," which are units of speech sounds that a person utters. For example, the spoken English word "cat" in the General American dialect comprises three phones [K], [AE], and [T], and the word "cot" comprises three phones [K], [AA], and [T]. In this example, "cat" and "cot" share the initial and final consonants but employ different vowels. Codebooks are structured to provide a one-to-one mapping between the phone entries in a source codebook and the phone entries in the target codebook.

U.S. Patent No. 5,327,521 describes a conventional voice conversion system using a codebook approach. An input signal from a source speaker is sampled and preprocessed by segmentation into "frames" corresponding to a speech unit. Each frame is matched to the "closest" source codebook entry and then mapped to the corresponding target codebook entry to obtain a phone in the voice of the target speaker. The mapped frames are concatenated to produce speech in the target voice. A disadvantage with this and similar conventional voice conversion systems is the introduction of artifacts at frame boundaries leading to a rather rough transition across target frames. Furthermore, the variation between the sound of the input speech frame and the closest matching source codebook entry is discarded, leading to a low quality voice conversion.

A common cause for the variation between the sounds in speech and in codebook is that sounds differ depending on their position in a word. For example, the /t/ phoneme has several "allophones." At the beginning of a word, as in the General American pronunciation of the word "top", the /t/ phoneme is an unvoiced, fortis, aspirated, alveolar stop. In an initial cluster with an /s/, as in the word "stop," it is an unvoiced, fortis, unaspirated, alveolar stop. In the middle of a word between vowels, as in "potter," it is an alveolar flap. At the end of a word, as in "pot," it is an unvoiced, lenis, unaspriated, alveolar stop. Although the allophones of a consonant like /t/ are pronounced differently, a codebook with only one entry for the /t/ phoneme will produce only one kind of /t/ sound and, hence, unconvincing output. Prosody also accounts for differences in sound, since a consonant or vowel will sound somewhat different when spoken at a higher or lower pitch, more or less rapidly, and with greater or lesser emphasis.

Accordingly, one conventional attempt to improve voice conversion quality is to greatly increase the amount of training data and the number of codebook entries to account for the different allophones of the same phoneme and different prosodic conditions. Greater codebook sizes lead to increased storage and computational costs. Conventional voice conversion systems also suffer in a loss of quality because they typically perform their codebook mapping in an acoustic space defined by linear predictive coding coefficients. Linear predictive coding is an all-pole modeling of speech and, hence, does not adequately represent the zeroes in a speech signal, which are more commonly found in nasal and sounds not originating at the glottis. Linear predictive coding also has difficulties with higher pitched sounds, for example, women's voices and children's voices.

SUMMARY OF THE INVENTION

There exists a need for a voice conversion system and methodology having improved quality output, but preferably still computationally tractable. Differences in sound due to word position and prosody need to be addressed without increasing the size of codebooks. Furthermore, there is a need to account for voice features that are not well supported by linear predictive coding, such as the glottal excitation, nasalized sounds, and sounds not originating at the glottis.

Accordingly, one aspect of the invention is a method and a computer-readable medium bearing instructions for transforming a source signal representing a source voice into a target signal representing a target voice. The source signal is preprocessed to produce a source signal segment, which is compared with source codebook entries to produce corresponding weights. The source signal segment is transformed into a target signal segment based on the weights and corresponding target codebook entries and post processed to generate the target signal. By computing a weighted average, a composite source voice can be mapped to a corresponding composite target voice, thereby reducing artifacts at frame boundaries and leading to smoother transitions between frame boundaries without having to employ a large number of codebook entries.

In another aspect of the invention, the source signal segment is compared with the source codebook entries as line spectral frequencies to facilitate the computation of the weighted average. In still another aspect of the invention, the weights are refined by a gradient descent analysis to further improve voice quality. In a further aspect of the invention, both vocal tract characteristics and excitation characteristics are transformed according to the weights, thereby handling excitation characteristics in a computationally tractable manner.

Additional needs, objects, advantages, and novel features of the present invention will be set forth in part in the description that follows, and in part, will become apparent upon examination or may be learned by practice of the invention. The objects and advantages of the invention may be realized and obtained by means of the instrumentalities and combinations particularly pointed out in the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

Fig. 1 schematically depicts a computer system that can implement the present invention;

Fig. 2 depicts codebook entries for a source speaker and a target speaker;

Fig 3 is a flowchart illustrating the operation of voice conversion according to an embodiment of the present invention;

Fig. 4 is a flowchart illustrating the operation of refining codebook weight by a gradient descent analysis according to an embodiment of the present invention; and

Fig 5 depicts a bandwidth reduction of formants of a weighted target voice spectrum according to an embodiment of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

A method and apparatus for voice conversion is described. In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention.

HARDWARE OVERVIEW

Figure 1 is a block diagram that illustrates a computer system 100 upon which an embodiment of the invention may be implemented. Computer system 100 includes a bus 102 or other communication mechanism for communicating information, and a processor (or a plurality of central processing units working in cooperation) 104 coupled with bus 102 for processing information. Computer system 100 also includes a main memory 106, such as a random access memory (RAM) or other dynamic storage device, coupled to bus 102 for storing information and instructions to be executed by processor 104. Main memory 106 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 104. Computer system 100 further includes a read only memory (ROM) 108 or other static storage device coupled to bus 102 for storing static information and instructions for processor 104. A storage device 110, such as a magnetic disk or optical disk, is provided and coupled to bus 102 for storing information and instructions.

Computer system 100 may be coupled via bus 102 to a display 111, such as a cathode ray tube (CRT), for displaying information to a computer user. An input device 113, including alphanumeric and other keys, is coupled to bus 102 for communicating information and command selections to processor 104. Another type of user input device is cursor control 115, such as a mouse, a trackball, or cursor direction keys for communicating direction information and command selections to processor 104 and for controlling cursor movement on display 111. This input device typically has two degrees of freedom in two axes, a first axis (e.g., *x*) and a second axis (e.g., *y*), that allows the device to specify positions in a plane. For audio output and input, computer system 100 may be coupled to a speaker 117 and a microphone 119, respectively.

The invention is related to the use of computer system 100 for voice conversion. According to one embodiment of the invention, voice conversion is provided by computer system 100 in response to processor 104 executing one or more sequences of one or more instructions contained in main memory 106. Such instructions may be read into main memory 106 from another computer-readable medium, such as storage device 110. Execution of the sequences of instructions contained in main memory 106 causes processor 104 to perform the process steps described herein. One or more processors in a multi-processing arrangement may also be employed to execute the sequences of instructions contained in main memory 106. In alternative embodiments, hard-wired circuitry may be used in place of or in combination with software instructions to implement the invention. Thus, embodiments of the invention are not limited to any specific combination of hardware circuitry and software.

The term "computer-readable medium" as used herein refers to any medium that participates in providing instructions to processor 104 for execution. Such a medium may take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Non-volatile media include, for example, optical or magnetic disks, such as storage device 110. Volatile media include dynamic memory, such as main memory 106. Transmission media include coaxial cables, copper wire and fiber optics, including the wires that comprise bus 102. Transmission media can also take the form of acoustic or light waves, such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, a PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave as described hereinafter, or any other medium from which a computer can read.

Various forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to processor 104 for execution. For example, the instructions may initially be borne on a magnetic disk of a remote computer. The remote computer can load the instructions into its dynamic memory and send the instructions over a telephone line using a modem. A modem local to computer system 100 can receive the data on the telephone line and use an infrared transmitter to convert the data to an infrared signal. An infrared detector coupled to bus 102 can receive the data carried in the infrared signal and place the data on bus 102. Bus 102 carries the data to main memory 106, from which processor 104 retrieves and executes the instructions. The instructions received by main memory 106 may optionally be stored on storage device 110 either before or after execution by processor 104.

Computer system 100 also includes a communication interface 120 coupled to bus 102. Communication interface 120 provides a two-way data communication coupling to a network link 121 that is connected to a local network 122. Examples of communication interface 120 include an integrated services digital network (ISDN) card, a modem to provide a data communication connection to a corresponding type of telephone line, and a local area network (LAN) card to provide a data communication connection to a compatible LAN. Wireless links may also be implemented. In any such implementation, communication interface 120 sends and receives electrical, electromagnetic or optical signals that carry digital data streams representing various types of information.

Network link 121 typically provides data communication through one or more networks to other data devices. For example, network link 121 may provide a connection through local network 122 to a host computer 124 or to data equipment operated by an Internet Service Provider (ISP) 126. ISP 126 in turn provides data communication services through the world wide packet data communication network, now commonly referred to as the "Internet" 128. Local network 122 and Internet 128 both use electrical, electromagnetic or optical signals that carry digital data streams. The signals through the various networks and the signals on network link 121 and through communication interface 120, which carry the digital data to and from computer system 100, are exemplary forms of carrier waves transporting the information.

Computer system 100 can send messages and receive data, including program code, through the network(s), network link 121, and communication interface 120. In the Internet example, a server 130 might transmit a requested code for an application program through Internet 128, ISP 126, local network 122 and communication interface 118. In accordance with the invention, one such downloaded application provides for voice conversion as described herein. The received code may be executed by processor 104 as it is received, and/or stored in storage device 110, or other non-volatile storage for later execution. In this manner, computer system 100 may obtain application code in the form of a carrier wave.

SOURCE AND TARGET CODEBOOKS

In accordance with the present invention, codebooks for the source voice and the target voice are prepared as a preliminary step, using processed samples of the source and target speech, respectively. The number of entries in the codebooks may vary from implementation to implementation and depends on a trade-off of conversion quality and computational tractability. For example, better conversion quality may be obtained by including a greater number of phones in various phonetic contexts but at the expense of increased utilization of computing resources and a larger demand on training data. Preferably, the codebooks include at least one entry for every phoneme in the conversion language. However, the codebooks may be augmented to include allophones of phonemes and common phoneme combinations may augment the codebook. Figure 2 depicts an exemplary codebook comprising 64 entries. Since vowel quality often depends on the length and stress of the vowel, a plurality of vowel phones for a particular vowel, for example, [AA], [AA1], and [AA2], are included in the exemplary codebook.

The entries in the source codebook and the target codebooks are obtained by recording the speech of the source speaker and the target speaker, respectively, and their speech into phones. According to one training approach, the source and target speakers are asked to utter words and sentences for which an orthographic transcription is prepared. The training speech is sampled at an appropriate frequency such as 16 kHz and automatically segmented using, for example, a forced alignment to a phonetic translation of the orthographic transcription within an HMM framework using Mel-cepstrum coefficients and delta coefficients as described in more detail in C. Wightman & D. Talkin, *The Aligner User's Manual*, Entropic Reseach Laboratory, Inc., Washington, D.C., 1994.

Preferably, the source and target vocal tract characteristics in the codebook entries are represented as line spectral frequencies (LSF). In contrast to conventional approaches using linear prediction coefficients (LPC) or formant frequencies, line spectral frequencies can be estimated quite reliably and have a fixed range useful for real-time digital signal processing implementation. The line spectral frequency values for the source and target codebooks can be obtained by first determining the linear predictive coefficients $\mathbf{a}_k$ for the sampled signal according to well-known techniques in the art. For example, specialized hardware, software executing on a general purpose computer or microprocessor, or a combination thereof, can ascertain the linear predictive coefficients by such techniques as square-root or Cholesky decomposition, Levinson-Durbin recursion, and lattice analysis introduced by Itakura and Saito. The linear predictive coefficients $\mathbf{a}_k$, which are recursively related to a sequence of partial correlation (PARCOR) coefficients, form an inverse filter polynomial, $A(z) = 1 - \sum_{k=1}^{P} \mathbf{a}_k z^{-k}$, which may be augmented with +1 and -1, to produce following polynomials, wherein the angles of the roots, $\mathbf{w}_k$, are the line spectral frequencies:

$$P(z) = (1 - z^{-1}) \prod_{k=1,3,5,...}^{P-1} (1 - 2\cos(\mathbf{w}_k z^{-1} + z^{-1})) \qquad (1)$$

$$Q(z) = (1 + z^{-1}) \prod_{k=2,4,6,...}^{P-1} (1 - 2\cos(\mathbf{w}_k z^{-1} + z^{-1})) \qquad (2)$$

Preferably, a plurality of samples are taken for each source and target codebook entry and averaged or otherwise processed, such as taking the median sample or the sample closest to the mean, to produce a source centroid vector $\mathbf{S}_i$ and target vector centroid $\mathbf{T}_i$, respectively, where $i \in 1..L$, and L is size of the codebook. Line spectral frequencies can be converted back into linear predictive coefficients by generating a sequence of coefficients via polynomial $P(z)$ and $Q(z)$ and, thence, the linear predictive coefficients $\mathbf{a}_k$.

Thus, the source codebook and the target codebook have corresponding entries containing speech samples derived respectively from the source speaker and the target speaker. Referring again to Figure 2, the light curves in each codebook entry represent the (male) source speaker's voice and the dark curves in each codebook entry represent the (female) target speaker's voice.

CONVERTING SPEECH

When the appropriate codebooks for the source and target speakers have been prepared, input speech in the source voice is transformed into the voice of the target speaker, according to one embodiment of the present invention, by performing the steps illustrated in Fig. 3. In step 300, the input speech is preprocessed to obtain an input speech frame. More specifically, the input speech is sampled at an appropriate frequency such as 16 kHz, and the DC bias is removed as by mean removal. The sampled signal is also windowed to produce the input speech frame $x(n) = w(n)s(n)$, where $w(n)$ is a data windowing function providing a raised cosine window, *e.g.* a Hamming window or a Hanning window, or other window such a rectangular window or a center-weighted window.

In step 302, the input speech frame is converted into line spectral frequency format. According to one embodiment of the present invention, a linear predictive coding analysis is first performed to determine the predication coefficients $\mathbf{a}_k$ for the input speech frame. The linear predictive coding analysis is of an appropriate order, for example, from an 14th order to a 30th order analysis, such as an 18th order or 20th order analysis. Based on the predication coefficients $\mathbf{a}_k$, a line spectral frequency vector $\mathbf{w}_k$ is derived, as by the use of polynomials $P(z)$ and $Q(z)$, explained in more detail herein above.

CODEBOOK WEIGHTS

Conventional voice conversions by codebook methodologies suffer from loss of information due to matching only to a single, "closest" source phone. Consequently, artifacts may be introduced at speech frame boundaries, leading to rough transitions from one frame to the next. Accordingly, one embodiment of the invention matches the incoming speech frame to a weighted average of a plurality of codebook entries rather than to a single codebook entry. The weighting of codebook entries preferably reflects perceptual criteria. Use of a plurality of codebook entries smoothes the transition between speech frames and captures the vocal nuances between related sounds in the target speech output. Thus, in step 304, codebook weights $\mathbf{v}_i$ are estimated by comparing the input line spectral frequency vector $\mathbf{w}_k$ with each centroid vector $\mathbf{S}_i$ in the source codebook to calculate a corresponding distance $\mathbf{d}_i$:

$$\mathbf{d}_i = \sum_{k=1}^{P} \mathbf{h}_k \mid \mathbf{w}_k - \mathbf{S}_{ik} \mid, i \in 1..L \qquad (3)$$

where L is the codebook size. The distance calculation includes a weight factor $\mathbf{h}_k$, which is based on a perceptual criterion wherein closely spaced line spectral frequency pairs, which are likely to correspond to formant locations, are assigned higher weights:

$$\mathbf{h}_k = \frac{e^{-0.05\,|K-k|}}{\min(|\mathbf{w}_k - \mathbf{w}_{k-1}|, |\mathbf{w}_k - \mathbf{w}_{k+1}|)}, k \in 1..P \qquad (4)$$

where K is 3 for voiced sounds and 6 for unvoiced, since the average energy decreases (for voiced sounds) and increases (for unvoiced sounds) with increasing frequency. Based on the calculated distances $\mathbf{d}_i$, the normalized codebook weights $\mathbf{v}_i$ are obtained as follows:

$$\mathbf{v}_i = \frac{e^{-\gamma \mathbf{d}_i}}{\sum_{l=1}^{L} e^{-\gamma \mathbf{d}_l}}, i \in 1..L \qquad (5)$$

where the value of γ for each frame is found by an incremental search in the range of 0.2 to 2.0 with the criterion of minimizing the perceptual weighted distance between the approximated line spectral frequency vector $\mathbf{v}\mathbf{S}_k$ and the input line spectral frequency vector $\mathbf{w}_k$.

CODEBOOK WEIGHT REFINEMENT

In some applications, even the normalized codebook weights $\mathbf{v}_i$ may not be an optimal set of weights that would represent the original speech spectrum. According to one embodiment of the present invention, a gradient descent analysis is performed to improve the estimated codebook weights $\mathbf{v}_i$. Referring to the flowchart illustrated in Fig. 4, one implementation of a gradient descent analysis comprises an initialization step 400 wherein an error value E is initialized to a very high number and a convergence constant η is initialized to a suitable value from 0.05 to 0.5 such as 0.1.

In the main loop of the gradient descent analysis, starting at step 402, an error vector **e** is calculated based on the distance between the approximated line spectral frequency vector v**S** and the input line spectral frequency vector **w** and weighted by the height factor **h**. In step 404, the error value E is saved in an old error variable *oldE* and new error value E is calculated from the error vector **e**, for example, by a sum of the absolute values or by a sum of squares. In step 406, the codebook weights $\mathbf{v}_i$ are updated by an addition of the error with respect to the source codebook vector **eS**, factored by the convergence constant η and constrained to be positive to prevent unrealistic estimates. In order to reduce computation according to one embodiment of the present invention, the convergence constant η is adjusted based on the reduction in error. Specifically, if there is a reduction in error, the convergence constant η is increased, otherwise it is decreased (step 408). The main loop is repeated until the reduction in error fall below an appropriate threshold, such as one part in ten thousand (step 410).

It is observed that only a few codebook entries are assigned significantly large weight values in the initial weight vector estimate **v**. Therefore, one embodiment of the present invention, in order to save computation resources, updates the weights **v** in step 406 only on the first few largest weights, *e.g.* on the five largest weights. Use of this gradient descent method has resulted in an additional 15% reduction in the average Itakura-Saito distance between the original spectra $\mathbf{w}_k$ and the approximated spectra $\mathbf{vS}_k$. The average spectral distortion (SD), which is a common spectral quantizer performance evaluation, was also reduced from 1.8 dB to 1.4 dB.

VOCAL TRACT SPECTRUM MAPPING

Referring back to Figure 3, in step 306, a target vocal tract filter $V_t(\omega)$ is calculated as a weighted average of the entries in the target codebook to represent the voice of the target speaker for the current speech frame. According to an embodiment of the present invention, the refined codebook weights $\mathbf{v}_i$ are applied to the target line spectral frequency vectors $\mathbf{T}_i$ to construct the target line spectral frequency vector $\mathbf{vT}_k$:

$$\tilde{\mathbf{w}}_k = \sum \mathbf{v}_i \mathbf{T}_{ik}, k \in 1..P \qquad (7)$$

The target line spectral frequencies are then converted into target linear prediction coefficients $\tilde{\mathbf{a}}_k$, for example by way of polynomials $P(z)$ and $Q(z)$. The target linear prediction coefficients $\tilde{\mathbf{a}}_k$ are in turn used to estimate the target vocal tract filter $V_t(\omega)$:

$$V_t(\omega) = \left| \frac{1}{1 - \sum_{k=1}^{P} \tilde{\mathbf{a}}_k e^{-jk\omega}} \right|^{\beta}, \qquad (8)$$

where β should theoretically be 0.5. The averaging of line spectral frequencies, however, often results in formants, or spectral peaks, with larger bandwidths, which is heard as a buzz artifact. One approach in addressing this problem is to increase the value of β, which adjusts the dynamic range of the spectrum and, hence, reduce the bandwidths of the formant frequencies. One disadvantage with increasing β, however, is that the bandwidth is reduced also in other frequency bands besides the formant locations, thereby warping the target voice spectrum.

Accordingly, another approach is to reduce the bandwidths of the formants by adjusting the line spectral frequencies directly. The target line spectrum pairs $\tilde{\mathbf{w}}_i^j$ and $\tilde{\mathbf{w}}_{i+1}^j$ around the first F formant frequency locations $\mathbf{f}_j, j \in 1..F$, are modified, wherein F is set to a small integer such as four (4). The source formant bandwidths $\mathbf{b}_j$ and the target formant bandwidths $\tilde{\mathbf{b}}_j$ are used to estimate a bandwidth adjustment ratio, r:

$$r = \frac{\sum_{j=1}^{F} \mathbf{b}_j}{\sum_{j=1}^{F} \tilde{\mathbf{b}}_j} \qquad (9)$$

Accordingly, each pair of target line spectrum $\tilde{\mathbf{w}}_i^j$ and $\tilde{\mathbf{w}}_{i+1}^j$ around corresponding formant frequency location $\mathbf{f}_j$ is adjusted as follows:

$$\tilde{\mathbf{w}}_i^j \leftarrow \tilde{\mathbf{w}}_i^j + (1-r)(\mathbf{f}_j - \tilde{\mathbf{w}}_i^j), j \in 1..F \qquad (10)$$

and

$$\tilde{\mathbf{w}}_{i+1}^{j} \leftarrow \tilde{\mathbf{w}}_{i+1}^{j} + (1-r)(\mathbf{f}_j - \tilde{\mathbf{w}}_{i+1}^{j}), j \in 1..F \tag{11}$$

A minimum bandwidth value, *e.g.* $\frac{\mathbf{f}_j}{20}$ Hz or 50Hz, may be set in order to prevent the estimation of unreasonable bandwidths. Fig. 5 illustrates a comparison of the target speech power spectrum for the [AA] vowel before (light curve 500) and after (dark curve 510) the application of this bandwidth reduction technique. Reduction in the bandwidth of the first four formants 520, 530, 540, and 550, results in higher and more distinct spectral peaks. According to detailed observations and subjective listening tests, use of this bandwidth reduction technique has resulted in improved voice output quality.

EXCITATION CHARACTERISTICS MAPPING

Another factor that influences speaker individuality and, hence, voice conversion quality is excitation characteristics. The excitation can be very different for different phonemes. For example, voiced sounds are excited by a periodic pulse train or "buzz," and unvoiced sounds are excited by white noise or "hiss." According to one embodiment of the present invention, the linear predictive coding residual is used as an approximation of the excitation signal. In particular, the linear predictive coding residuals for each entry in the source codebook and the target codebook are collected as the excitation signals from the training data to compute a corresponding short-time average discrete Fourier analysis or pitch-synchronous magnitude spectrum of the excitation signals. The excitation spectra are used to formulate excitation transformation spectra for entries of the source codebook, $\mathbf{U}_i^s(\omega)$, and the target codebook, $\mathbf{U}_i^t(\omega)$. Since linear predictive coding is an all-pole model, the formulated excitation transformation filters serve to transform the zeros in the spectrum as well, thereby further improving the quality of the voice conversion.

Referring back to Figure 3, in step 308, the excitations in the input speech segment are transformed from the source voice to the target voice by the same codebook weights $\mathbf{v}_i$ used in transforming the vocal tract characteristics. Specifically, an overall excitation filter is constructed as a weighted combination of the excitation codebook excitation spectra:

$$H_g(\omega) = \sum \mathrm{v}_i \frac{\mathbf{U}_i^t(\omega)}{\mathbf{U}_i^s(\omega)} \qquad (12)$$

According to one embodiment of the present invention, the overall excitation filter $H_g(\omega)$ is applied to the linear predictive coding residual $e(n)$ of the input speech signal $x(n)$ to produce a target excitation filter:

$$G_t(\omega) = H_g(\omega)\,\mathrm{DFT}\{e(n)\} \qquad (13)$$

where the linear predictive coding residual $e(n)$ is given by:

$$e(n) = x(n) - \sum_{k=1}^{P} \mathbf{a}_k x(n-k) \qquad (14)$$

Both the vocal tract characteristics and the excitations characteristics are transformed in the same computational framework, by computing a weighted average of codebook entries. Accordingly, this aspect of the present invention enables the incorporation of excitation characteristics within a voice conversion system in a computationally tractable manner.

TARGET SPEECH FILTER

Referring again to Fig. 3, in step 310, a target speech filter $Y(\omega)$ is on the basis of the vocal tract filter $V_t(\omega)$ and, in some embodiments of the present invention, the excitation filter $G_t(\omega)$. According to one embodiment, target speech filter $Y(\omega)$ is defined as the the excitation filter $G_t(\omega)$ followed by the vocal tract filter $V_t(\omega)$:

$$Y(\omega) = G_t(\omega)V_t(\omega). \qquad (15)$$

In accordance with another embodiment of the present invention, further refinement to the construction of the target speech filter $Y(\omega)$ may be desirable for improved handling of unvoiced sounds. The incoming speech spectrum $X(\omega)$, derived from the sampled and windowed input speech $x(n)$, can be represented as

$$X(\omega) = G_s(\omega)V_s(\omega), \qquad (16)$$

where $G_s(\omega)$ and $V_s(\omega)$ represent the source speaker excitation and vocal tract spectrum filters, respectively. Consequently, the target speech spectrum filter $Y(\omega)$ can be formulated as:

$$Y(\omega)=\left[\frac{G_t(\omega)}{G_s(\omega)}\right]\left[\frac{V_t(\omega)}{V_s(\omega)}\right]X(\omega) \tag{17}$$

Using the overall excitation filter $H_g(\omega)$ as an estimate of the excitation filter, the target speech spectrum filter $Y(\omega)$ becomes:

$$Y(\omega)=H_g(\omega)\left[\frac{V_t(\omega)}{V_s(\omega)}\right]X(\omega) \tag{18}$$

When the amount of the training data is small or when the accuracy of the segmentation in question, unvoiced segments are difficult to represent accurately, thereby leading to a mismatch in the source and target vocal tract filters. Accordingly, one embodiment of the present invention, estimates a source speaker vocal tract spectrum filter $V_s(\omega)$ differently for voiced segments and for unvoiced segments. For voiced segments, the source speaker vocal tract spectrum filter $V_s(\omega)$ is replaced with the spectrum derived from the original linear predictive coefficient vector $\mathbf{a}_k$:

$$V_s(\omega)=\frac{1}{1-\sum_{k=1}^{P}\mathbf{a}_k e^{-jk\omega}}. \tag{19}$$

On the other hand, the linear predictive vector approximation coefficients, derived from the codebook weighted line spectral frequency vector approximation $\mathbf{vS}_k$, is used to determine the source speaker vocal tract spectrum filter $V_s(\omega)$ for unvoiced segments.

In step 312, the result of applying $Y(\omega)$ for the current segment is post processed into a time-domain target signal in the voice of the target speaker. More specifically, an inverse discrete Fourier transform is applied to produce the synthetic target voice:

$$y(n)=\mathrm{Re}\{\mathrm{IDFT}\{Y(\omega)\}\}. \tag{20}$$

PROSODY TRANSFORMATION

According to one embodiment of the present invention, prosodic transformations may be applied to the frequency domain target voice signal $Y(\omega)$ before post processing into the time domain. Prosodic transformations allow the target voice to match the source voice in pitch, duration, and stress. For example, a pitch scale modification factor β at each frame can be set as

$$\beta = \frac{\sqrt{\frac{\sigma_t^2}{\sigma_s^2}}(f_0 - \mu_s) + \mu_t}{f_0}, \tag{21}$$

where σ_s^2 is the source pitch variance, σ_t^2 is the target pitch variance, f_0 is the source speaker fundamental frequency, μ_s is the source mean pitch value, and μ_t is the target mean pitch value. For duration characteristics, a time-scale modification factor γ can be set according to the same codebook weights:

$$\gamma = \sum_{i=1}^{L} \mathbf{v}_i \frac{d_i^t}{d_i^s}, \tag{22}$$

where d_i^s is the average source speaker duration and d_i^t is the average target speaker duration. For the speakers' stress characteristics, an energy-scale modification factor η can be set according to the same codebook weights:

$$\eta = \sum_{i=1}^{L} \mathbf{v}_i \frac{e_i^t}{e_i^s}, \tag{23}$$

where e_i^s is the average source speaker RMS energy and e_i^t is the average target speaker RMS energy.

The pitch-scale modification factor β, the time-scale modification factor γ, and the energy scaling factor η are applied by an appropriate methodology, such as within a pitch-synchronous overlap-add synthesis framework, to perform the prosodic synthesis. One overlap-add synthesis methodology is explained in more detail in the commonly assigned Application Ser. No. __________, entitled "Prosody Modification System and Methodology," filed concurrently by Francisco M. Gimenez de los Galenes, the contents of which are herein incorporated by reference.

While this invention has been described in connection with what is presently considered to be the most practical and preferred embodiment, it is to be understood that the invention is not limited to the disclosed embodiment, but on the contrary, is intended to cover various modifications and equivalent arrangements included within the spirit and scope of the appended claims.

CLAIMS

WHAT IS CLAIMED IS:

1. A method of transforming a source signal representing a source voice into a target signal representing a target voice, said method comprising the machine-implemented steps of:

preprocessing said source signal to produce a source signal segment;

comparing the source signal segment with a plurality of source codebook entries representing speech units in said source voice to produce therefrom a plurality of corresponding weights;

transforming the source signal segment into a target signal segment based on the plurality of weights and a plurality of target codebook entries representing speech units in said target voice, said target codebook entries corresponding to the plurality of source codebook entries; and

post processing the target signal segment to generate said target signal.

2. A method as in claim 1, wherein the step of preprocessing said source signal includes the step of sampling said source signal to produce a sampled source signal.

3. A method as in claim 2, wherein the step of preprocessing said source signal includes the step of segmenting said sampled source signal to produce the source signal segment.

4. A method as in claim 1, wherein the step of comparing the source signal segment to produce therefrom a plurality of corresponding weights includes the step of comparing

the source signal segment to produce therefrom a plurality of corresponding perceptual weights.

5. A method as in claim 1, wherein the step of comparing the source signal segment includes the steps of:

converting the source signal segment into a plurality of line spectral frequencies; and

comparing the plurality of line spectral frequencies with the plurality of the source code entries to produce therefrom the plurality of the respective weights, wherein each of the source code entries include a respective plurality of line spectral frequencies.

6. A method as in claim 5, wherein the step of converting the source signal segment includes the steps of:

determining a plurality of coefficients for the source signal segment; and

converting the plurality of coefficients into the plurality of line spectral frequencies.

7. A method as in claim 6, wherein the step of determining a plurality of coefficients includes the step of determining a plurality of linear prediction coefficients or PARCOR coefficients.

8. A method as in claim 5, wherein the step of comparing the plurality of line spectral frequencies includes the steps of:

computing a plurality of distances between the source signal segment, represented by the plurality of line spectral frequencies, and each of the plurality of the respective source code entries, represented by a respective plurality of line spectral frequencies; and

producing the plurality of the weights based on the plurality of respective distances.

9. A method as in claim 8, further including the step of refining the plurality of weights by a gradient descent method.

10. A method as in claim 1, wherein the step of transforming the source signal segment into a target signal segment based on the plurality of weights and a plurality of target codebook entries includes the step of transforming vocal tract characteristics of the source signal segment into the target signal segment based on the plurality of weights and a plurality of target codebook entries.

11. A method as in claim 10, wherein the step of transforming vocal tract characteristics includes the step of reducing formant bandwidths in the target signal segment.

12. A method as in claim 10, wherein the step of transforming the source signal segment into a target signal segment based on the plurality of weights and a plurality of target codebook entries includes the step of transforming excitation characteristics of the source signal segment into the target signal segment based on the plurality of weights.

13. A method as in claim 1, further including the step of modifying the prosody of the target signal segment based on the plurality of weights.

14. A method as in claim 13, wherein the step of modifying the prosody of the target signal segment based on the plurality of weights includes the step of modifying the duration of the target signal segment.

15. A method as in claim 13, wherein the step of modifying the prosody of the target signal segment based on the plurality of weights includes the step of modifying the stress of the target signal segment.

16. A computer-readable medium bearing instructions for transforming a source signal representing a source voice into a target signal representing a target voice, said instructions arranged, when executed, to cause one or more processors to perform the steps of:

preprocessing said source signal to produce a source signal segment;

comparing the source signal segment with a plurality of source codebook entries representing speech units in said source voice to produce therefrom a plurality of corresponding weights;

transforming the source signal segment into a target signal segment based on the plurality of weights and a plurality of target codebook entries representing speech units in said target voice, said target codebook entries corresponding to the plurality of source codebook entries; and

post processing the target signal segment to generate said target signal.

17. A computer-readable medium as in claim 16, wherein the step of preprocessing said source signal includes the step of sampling said source signal to produce a sampled source signal.

18. A computer-readable medium as in claim 17, wherein the step of preprocessing said source signal includes the step of segmenting said sampled source signal to produce the source signal segment.

19. A method as in claim 16, wherein the step of comparing the source signal segment to produce therefrom a plurality of corresponding weights includes the step of comparing the source signal segment to produce therefrom a plurality of corresponding perceptual weights.

20. A computer-readable medium as in claim 16, wherein the step of comparing the source signal segment includes the steps of:

converting the source signal segment into a plurality of line spectral frequencies; and

comparing the plurality of line spectral frequencies with the plurality of the source code entries to produce therefrom the plurality of the respective weights, wherein each of the source code entries include a respective plurality of line spectral frequencies.

21. A computer-readable medium as in claim 20, wherein the step of converting the source signal segment includes the steps of:

determining a plurality of coefficients for the source signal segment; and

converting the plurality of coefficients into the plurality of line spectral frequencies.

22. A computer-readable medium as in claim 21, wherein the step of determining a plurality of coefficients includes the step of determining a plurality of linear prediction coefficients or PARCOR coefficients.

23. A computer-readable medium as in claim 20, wherein the step of comparing the plurality of line spectral frequencies includes the steps of:

computing a plurality of distances between the source signal segment, represented by the plurality of line spectral frequencies, and each of the plurality of the respective source code entries, represented by a respective plurality of line spectral frequencies; and

producing the plurality of the weights based on the plurality of respective distances.

24. A computer-readable medium as in claim 23, further including the step of refining the plurality of the weight by a gradient descent method.

25. A computer-readable medium as in claim 16, wherein the step of transforming the source signal segment into a target signal segment based on the plurality of weights and a plurality of target codebook entries includes the step of transforming vocal tract characteristics of the source signal segment into the target signal segment based on the plurality of weights and a plurality of target codebook entries.

26. A computer-readable medium as in claim 25, wherein the step of transforming vocal tract characteristics includes the step of reducing formant bandwidths in the target signal segment.

27. A computer-readable medium as in claim 25, wherein the step of transforming the source signal segment into a target signal segment based on the plurality of weights and a plurality of target codebook entries includes the step of transforming excitation characteristics of the source signal segment into the target signal segment based on the plurality of weights.

28. A computer-readable medium as in claim 16, wherein the instructions, when executed, are further arranged to perform the step of modifying the prosody of the target signal segment based on the plurality of weights.

29. A computer-readable medium as in claim 28, wherein the step of modifying the prosody of the target signal segment based on the plurality of weights includes the step of modifying the duration of the target signal segment.

30. A computer-readable medium as in claim 28, wherein the step of modifying the prosody of the target signal segment based on the plurality of weights includes the step of modifying the stress of the target signal segment.

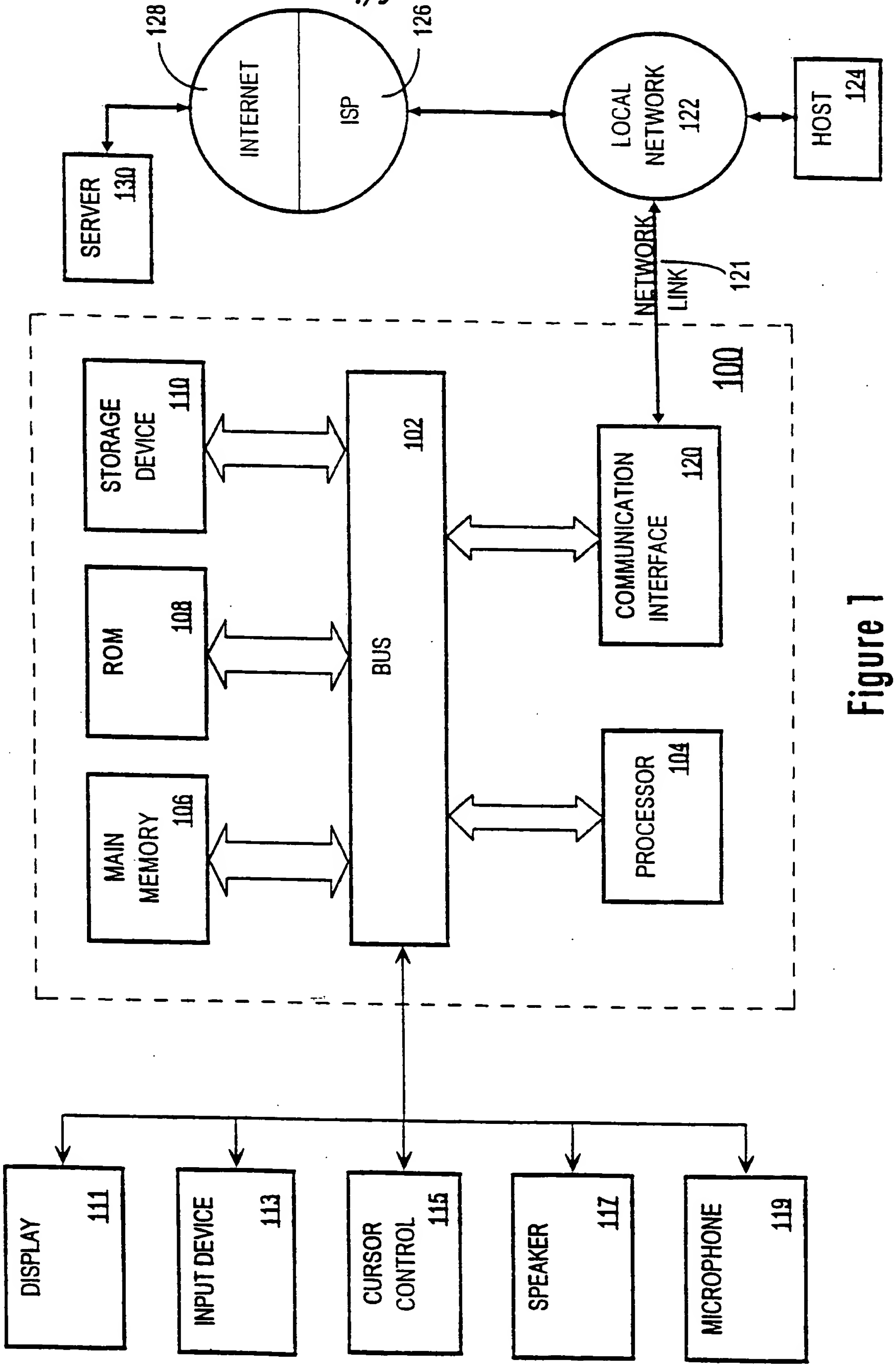

SERVER 130
INTERNET
ISP
128
126
LOCAL NETWORK 122
HOST 124
NETWORK LINK
121
100
STORAGE DEVICE 110
ROM 108
MAIN MEMORY 106
BUS 102
COMMUNICATION INTERFACE 120
PROCESSOR 104
DISPLAY 111
INPUT DEVICE 113
CURSOR CONTROL 115
SPEAKER 117
MICROPHONE 119

Figure 1

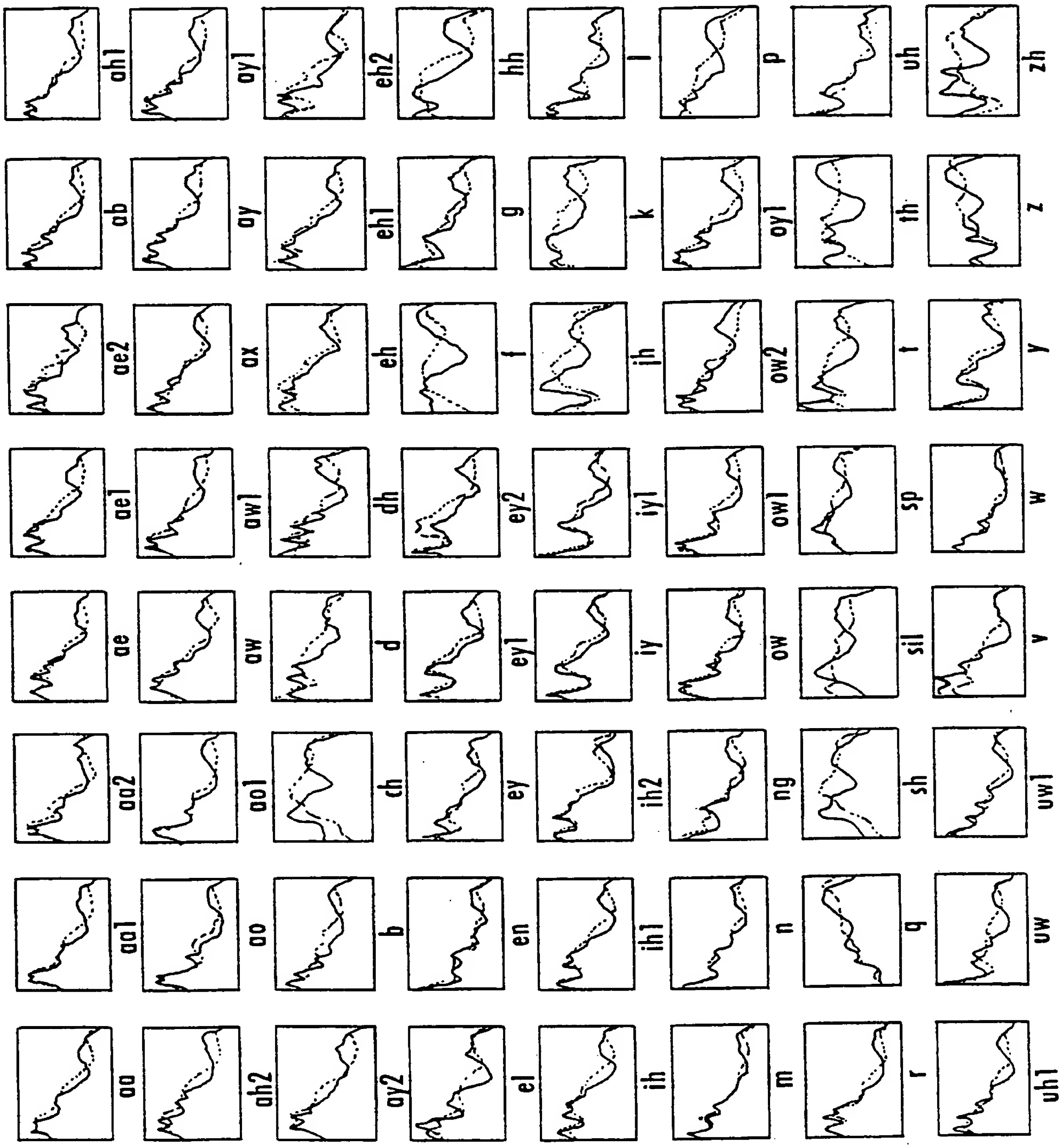

aa
aa1
aa2
ae
ae1
ae2
ab
ah1
ah2
ao
ao1
aw
aw1
ax
ay
ay1
ay2
b
ch
d
dh
eh
eh1
eh2
el
en
ey
ey1
ey2
f
g
hh
ih
ih1
ih2
iy
iy1
jh
k
l
m
n
ng
ow
ow1
ow2
oy1
p
r
q
sh
sil
sp
t
th
uh
uh1
uw
uw1
v
w
y
z
zh

Figure 2

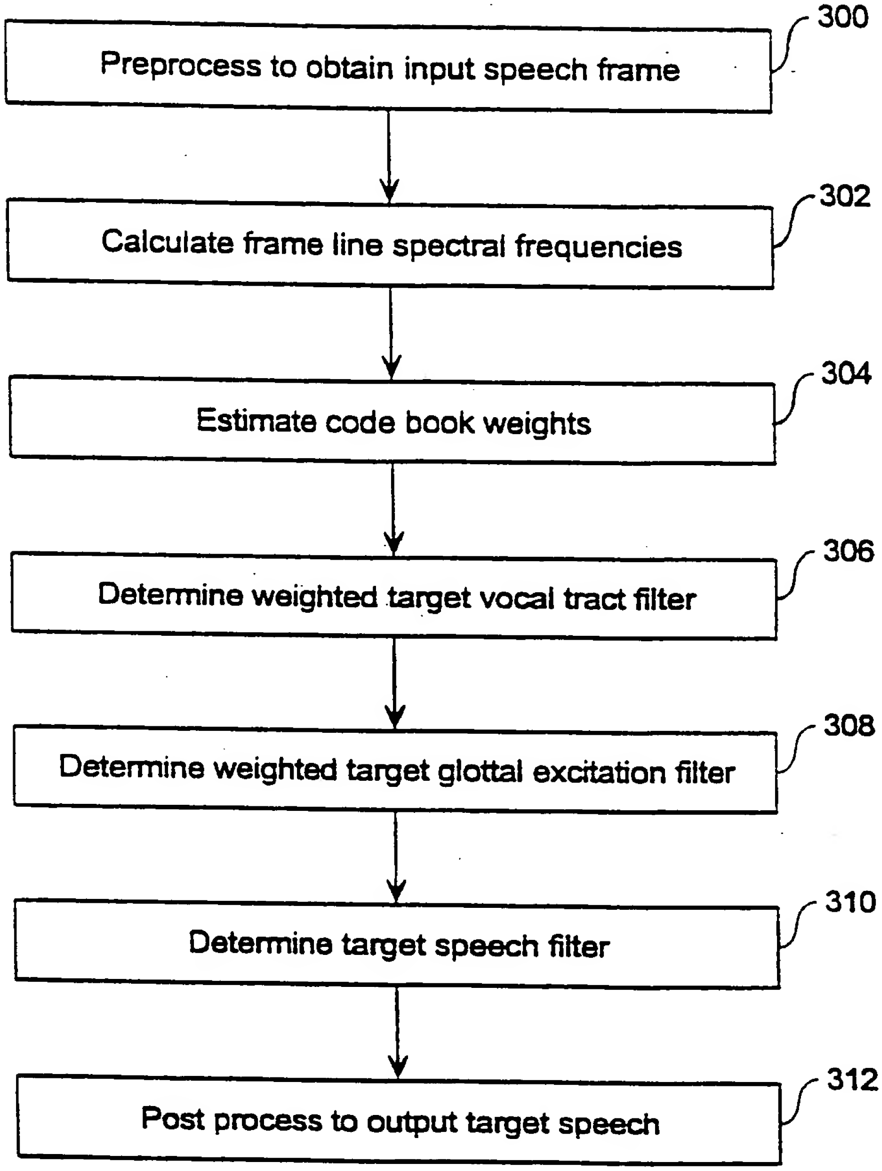

300
Preprocess to obtain input speech frame
302
Calculate frame line spectral frequencies
304
Estimate code book weights
306
Determine weighted target vocal tract filter
308
Determine weighted target glottal excitation filter
310
Determine target speech filter
312
Post process to output target speech

Figure 3

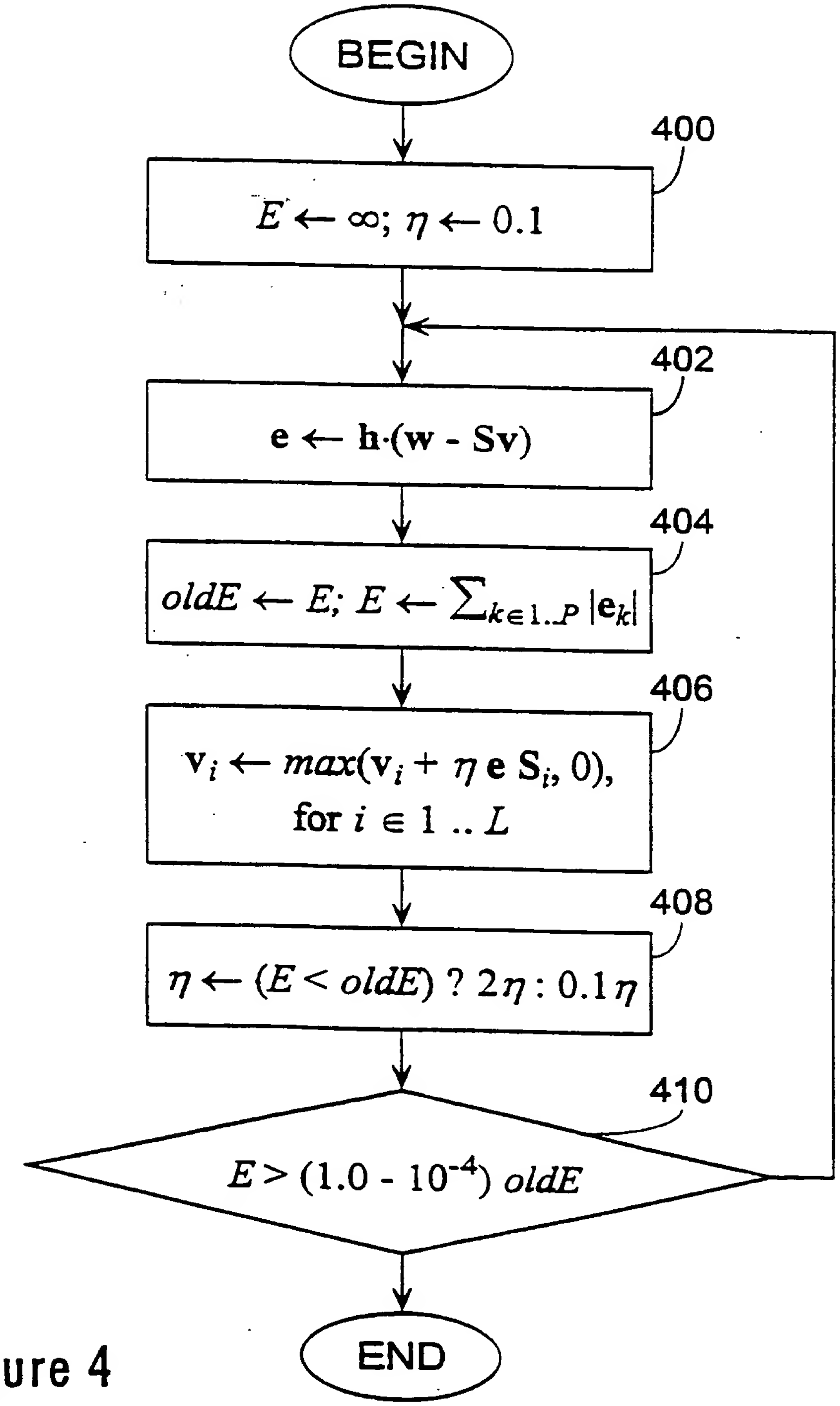

BEGIN
400
$E \leftarrow \infty; \eta \leftarrow 0.1$
402
$\mathbf{e} \leftarrow \mathbf{h} \cdot (\mathbf{w} - \mathbf{S}\mathbf{v})$
404
$oldE \leftarrow E; E \leftarrow \sum_{k \in 1..P} |\mathbf{e}_k|$
406
$\mathbf{v}_i \leftarrow max(\mathbf{v}_i + \eta\, \mathbf{e}\, \mathbf{S}_i, 0)$, for $i \in 1 .. L$
408
$\eta \leftarrow (E < oldE)\ ?\ 2\eta : 0.1\eta$
410
$E > (1.0 - 10^{-4})\ oldE$
END

Figure 4

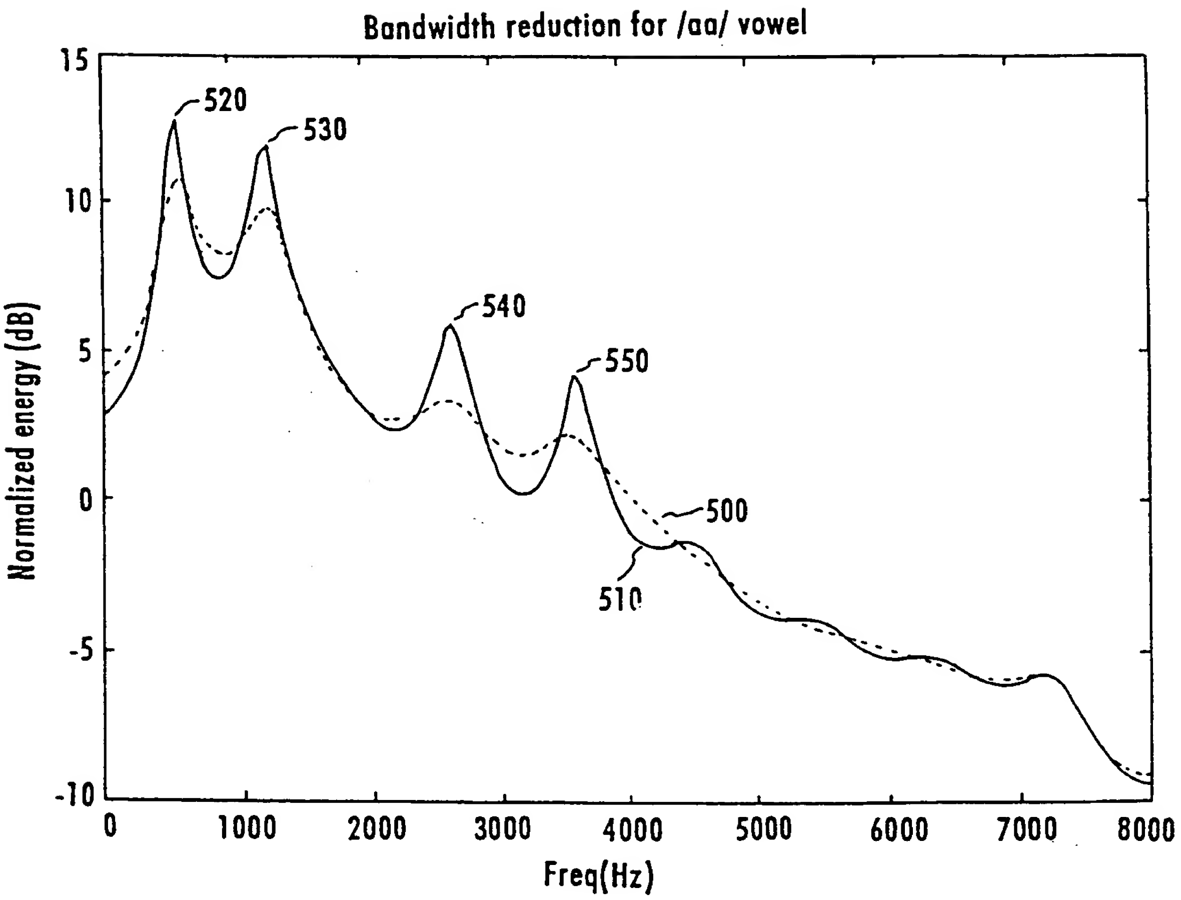

Bandwidth reduction for /aa/ vowel
15
10
5
0
-5
-10
Normalized energy (dB)
0
1000
2000
3000
4000
5000
6000
7000
8000
Freq(Hz)
520
530
540
550
500
510

Figure 5

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G10L 3/02

A3

(11) International Publication Number: WO 98/35340

(43) International Publication Date: 13 August 1998 (13.08.98)

(21) International Application Number: PCT/US98/01538

(22) International Filing Date: 27 January 1998 (27.01.98)

(30) Priority Data:
60/036,227 27 January 1997 (27.01.97) US

(71) Applicant *(for all designated States except US)*: ENTROPIC RESEARCH LABORATORY, INC. [US/US]; Suite G100, 400 North Capitol Street, N.W., Washington, DC 20001 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: ARSLAN, Levent, M. [TR/US]; 850 Randolph Street #811, Arlington, VA 22203 (US). TALKIN, David [US/US]; 1727 Lansing Court, McLean, VI 22101 (US).

(74) Agents: CARLSON, Stephen, C. et al.; Lowe Price Leblanc & Becker, Suite 300, 99 Canal Center Plaza, Alexandria, VA 22314 (US).

(81) Designated States: AU, CA, IL, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

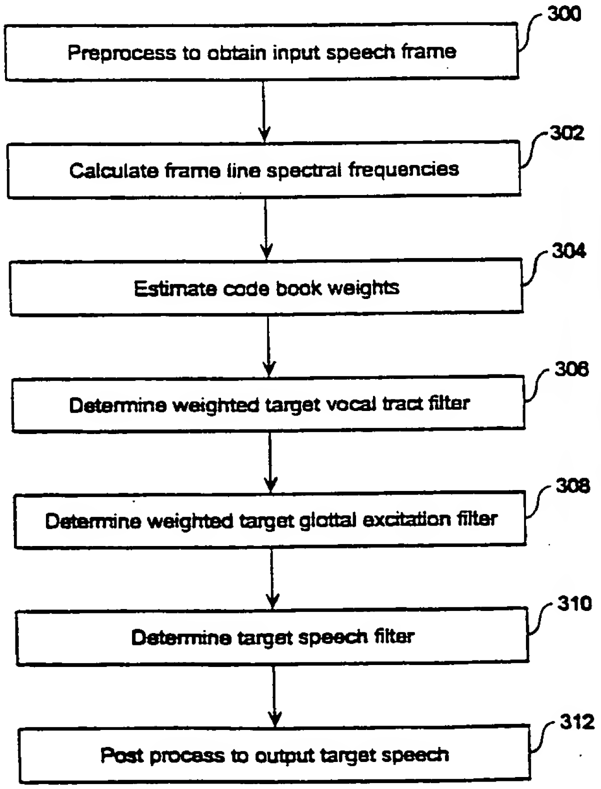
Published
With international search report.

(88) Date of publication of the international search report:
19 November 1998 (19.11.98)

(54) Title: VOICE CONVERSION SYSTEM AND METHODOLOGY

(57) Abstract

A voice conversion system and methodology employing a codebook mapping approach to transforming a source voice to sound like a target voice. Each speech frame is represented by a weighted average of codebook entries (304). The weights represent a perceptual distance of the speech frame and may be refined by a gradient descent analysis. The vocal tract characteristics, represented by a line spectral frequency vector (302), the excitation characteristics (308), represented by a linear predictive coding residual, the duration, and the amplitude of the speech frame are transformed in the same weighted-average framework.


300
Preprocess to obtain input speech frame
302
Calculate frame line spectral frequencies
304
Estimate code book weights
306
Determine weighted target vocal tract filter
308
Determine weighted target glottal excitation filter
310
Determine target speech filter
312
Post process to output target speech

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/01538

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G10L 3/02
US CL : 704/270
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 704/270, 269, 258, 270, 272, 263, 266

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, IEEE
search terms: speech, voice, weight?, synthesi?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,327,521 A (SAVIC ET AL) 05 July 1994 (05.07.94) column 2, lines 1-55, column 4 line 50.	1-30
Y,P	US 5,704,006 A (IWAHASHI) 30 December 1997 (30.12.97) column 2, lines 33-60, column 6 lines 1-20, column 6 lines 60-64, column 9 line 40.	1-30

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

16 APRIL 1998

Date of mailing of the international search report

19 AUG 1998

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
HAROLD ZINTEL
Telephone No. (703) 305-2381